

# 基于 EM 算法的男女生比例预测问题

SY2303129 赵秋驰

2023 年 11 月 28 日

## 摘要

针对性别比例未知、但身高数据已知的班级男女生比例预测问题，本文基于贝叶斯公式、全概率公式等数学工具，假定男女生身高各自服从于不同参数的正态分布，通过建立 EM 模型较好地解决了该预测问题。

## 1 准备知识

### 1.1 正态分布

假设男女生身高服从于不同参数的正态分布。设学生身高为  $x$ ，期望值为  $\mu$ ，方差为  $\sigma$ 。则有：

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

### 1.2 贝叶斯公式

假设学生中男性占比为  $p_1$ ，女性占比为  $p_2$ 。给定某个身高样本，由贝叶斯公式可知，该样本属于男生的后验概率  $p_{i1}$  为：

$$p_{i1} = \frac{p_1 N(x_i; \mu_1, \sigma_1)}{p_1 N(x_i; \mu_1, \sigma_1) + p_2 N(x_i; \mu_2, \sigma_2)} \quad (2)$$

属于女生的后验概率  $p_{i2}$  为：

$$p_{i2} = \frac{p_2 N(x_i; \mu_2, \sigma_2)}{p_1 N(x_i; \mu_1, \sigma_1) + p_2 N(x_i; \mu_2, \sigma_2)} \quad (3)$$

### 1.3 琴生不等式

琴生不等式为我们指出，对于一个凸函数  $f(x)$ ，有：

$$E(f(x)) \geq f(E(x)) \quad (4)$$

## 2 高斯混合模型推导

对于包含  $N$  个样本的班级身高数据，抽到身高样本为  $x_i$  的概率为：

$$P(x_i) = p_1 N(x_i; \mu_1, \sigma_1) + p_2 N(x_i; \mu_2, \sigma_2) \quad (5)$$

其似然函数为：

$$L(X) = \prod_{i=1}^N P(x_i) \quad (6)$$

根据琴生不等式，可以构造一个易于求导的下界似然函数来取代原函数：

$$L(X) = \sum_{i=1}^N \left( p_{i1} \ln \frac{p_1 N(x_i; \mu_1, \sigma_1)}{p_{i1}} + p_{i2} \ln \frac{p_2 N(x_i; \mu_2, \sigma_2)}{p_{i2}} \right) \quad (7)$$

## 3 EM 模型

### 3.1 EM 模型简介

EM 模型是一种基于概率分布的聚类优化模型。所谓 EM，就是指期望 (expectation) —— 优化 (maximization)。它引入隐变量，通过多次迭代收敛的方式来解决一些缺失数据的概率模型。

### 3.2 计算过程

#### 3.2.1 参数初始化

初始化男女生比例  $p_1$ 、 $p_2$ ，身高平均值  $\mu_1$ 、 $\mu_2$ ，身高方差  $\sigma_1$ 、 $\sigma_2$ 。  
以下推导均以男生为例，女生同理。

#### 3.2.2 优化步

首先进行性别比估计。

令似然函数 (7) 对  $p_1$  求偏导：

$$\frac{\partial L(p, \mu, \sigma)}{\partial p_1} = \sum_{i=1}^N \frac{1}{p_1} p_{i1} \quad (8)$$

由全概率公式，注意到：

$$p_1 + p_2 = 1 \quad (9)$$

式 (9) 是对式 (8) 的约束条件。为了解决 (8) 的优化问题，我们构造一个拉格朗日方程：

$$\frac{\partial L(p, \mu, \sigma)}{\partial p_1} = \sum_{i=1}^N \frac{1}{p_1} p_{i1} + \frac{\partial \lambda(p_1 + p_2 - 1)}{\partial p_1} \quad (10)$$

$$\sum_{i=1}^N \frac{1}{p_1} p_{i1} + \lambda = 0 \quad (11)$$

$$\lambda = -N \quad (12)$$

$$p_1 = \frac{1}{N} \sum p_{i1} \quad (13)$$

然后估计正态分布的两个参数——期望和方差。

依据似然函数对二者分别求偏导即可：

$$\frac{\partial L(p, \mu, \sigma)}{\partial \mu_1} = \sum_{i=1}^N (p_{i1} \frac{x_i - \mu_1}{\sigma_1^2}) \quad (14)$$

令偏导取 0：

$$\mu_1 = \frac{\sum_{i=1}^N p_{i1} x_i}{\sum_{i=1}^N p_{i1}} \quad (15)$$

同理求解方差值：

$$\frac{\partial L(p, \mu, \sigma)}{\partial \sigma_1} = \sum_{i=1}^N p_{i1} \left( -\frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3} \right) \quad (16)$$

解得方差值为：

$$\sigma_1 = \frac{\sum_{i=1}^N p_{i1} (x_i - \mu_1)^2}{\sum_{i=1}^N p_{i1}} \quad (17)$$

### 3.2.3 迭代

根据优化步计算出的男女生性别比、身高的期望和方差三个值，带回到式 (2)，计算出新的男女生性别比、身高的期望和方差。

由此循环往复地进行期望步——优化步迭代，直到参数收敛到一定的容许范围为止。

# 北京航空航天大学

BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

对于每个样本  $x_i \in \{x_1, \dots, x_N\}$ ,  $x_i$  所属的类别, 也就是样本类别未知,

假设每个类别的样本  $x$  各自服从高斯分布

$$\text{即 } P(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

为了使每个样本所属类别  $z_i$  尽量接近实际值, 可定义其似然函数为

$$L(\theta) = \prod_{i=1}^N P(x_i; \theta) \quad \text{其中 } \theta \text{ 包括 } \mu, \sigma$$

两边取对数得

$$\ln L(\theta) = \ln \left( \prod_{i=1}^N P(x_i; \theta) \right) = \sum_{i=1}^N \ln P(x_i; \theta)$$

$$\text{其中 } P(x_i; \theta) = \sum_{z_i} P(x_i, z_i; \theta)$$

但是, 直接最大化  $L(\theta)$  是很困难的, 所以需要取  $L(\theta)$  的一个下界, 通过优化这个下界来优化  $L(\theta)$

设  $Q_i$  为概率密度函数  $\sum_{z_i} Q_i(z_i) = 1 \quad Q_i(z_i) \geq 0$

$$\text{则有 } \sum_{i=1}^N \ln P(x_i; \theta) = \sum_{i=1}^N \ln \sum_{z_i} P(x_i, z_i; \theta)$$

$$= \sum_{i=1}^N \ln \sum_{z_i} Q_i(z_i) \frac{P(x_i, z_i; \theta)}{Q_i(z_i)}$$

$$\geq \sum_{i=1}^N \sum_{z_i} Q_i(z_i) \ln \frac{P(x_i, z_i; \theta)}{Q_i(z_i)}$$

上式成立的事件是:  $\frac{P(x_i, z_i; \theta)}{Q_i(z_i)} = c$   $c$  为固定常数

$$\text{那么有: } \frac{\sum_{z_i} P(x_i, z_i; \theta)}{\sum_{z_i} Q_i(z_i)} = c$$

$$\text{由于 } \sum_{z_i} Q_i(z_i) = 1$$

$$\text{所以 } Q_i(z_i) = \frac{P(x_i, z_i; \theta)}{c} = \frac{P(x_i, z_i; \theta)}{\sum_{z_i} P(x_i, z_i; \theta)} = P(z_i | x_i; \theta)$$

中国 · 北京 100191

37XUEYUANROADBEIJING 100191CHINA

图 1: 手写推导 1

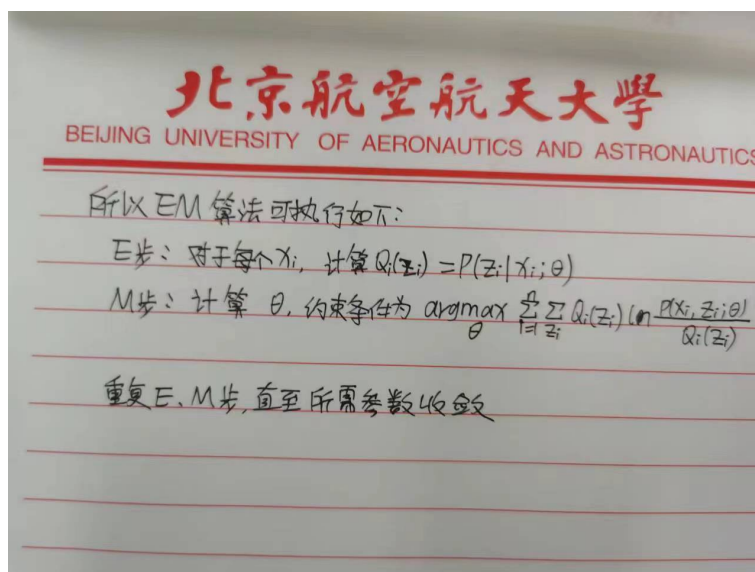


图 2: 手写推导 2

## 4 Python 实现与验证

根据前文的准备工作, 在 Python 中进行代码实现即可。

最终得到的预测结果为: 男生身高期望为 178.03cm, 标准差为 5.47, 男生占比 0.88; 女生身高期望为 163.83cm, 标准差为 2.46cm, 女生占比为 0.12。

表 1: EM 算法预测结果与实际情况

类别	EM 算法预测结果	真实值	误差
男生身高期望 /cm	178.03	178.33	0.17%
男生身高标准差 /cm	5.47	5.57	1.79%
女生身高期望 /cm	163.83	167.35	2.10%
女生身高标准差 /cm	2.46	5.51	55.35%
男生占比	88%	82%	7.31%
女生占比	12%	18%	33.33%

从结果来看, 由于原始数据样本量太小, 混合分布特性不明显, 更接近单正态分布。算法对男生身高的评估更准确, 但对女生身高情况及男女生比例评估误差较大。

## 5 模型拓展

在更复杂的情况下，比如每一个样本包含多个维度的特征时，需要对模型做一定的调整。

假设现在不仅给定了身高数据，还附带了对应的体重和年龄数据。也就是说每一个  $x_i$  都是 3 维的，那么正态分布函数应做如下修正：

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{(2\pi)^3 \det(\sigma)}} \exp^{-\frac{1}{2}(x-\mu)\sigma^{-1}(x-\mu)^T} \quad (18)$$

其中  $x_i$  为  $1 \times 3$  的向量， $\mu$  为  $1 \times 3$  的向量， $\sigma$  为  $3 \times 3$  的矩阵。

不难看出，与单变量的正态分布相比，多变量正态分布以协方差代替方差。实际上，方差是协方差的一种特殊情况。

## 6 总结

本文基于 EM 算法处理，实现了基于男女生身高数据的性比例预测问题。

## 参考文献

- [1] 左飞. 机器学习原理与实践: Python 版 [M]. 北京市: 清华大学出版社, 2021.
- [2] 魏宗舒. 概率论与数理统计教程 [M]. 北京市: 高等教育出版社, 2020: 80-87, 171-180
- [3] [印]M.Gopal. 机器学习及其应用 [M]. 黄智渊, 杨武兵, 等, 译: 北京: 机械工业出版社, 2020: 255-272